Post Exam 2 Extension - Hypothesis Testing

You have been studying regions of the chromosome of *Streptococcus agalactiae* that have unusual dinucleotide biases and have noticed that genes in those areas seem frequently to be annotated "hypothetical". This would be interesting, because you know that viral genes in general have not been well studied and their functions are often unknown and given the annotation "hypothetical". It makes sense to you that the unusual regions contain viral sequences.

There are at least two problems with this idea:

1. Maybe you're being fooled. You haven't looked at a lot of genes in your young life and – who knows – maybe the high incidence of hypothetical genes is common throughout the genome.
2. Another source of hypothetical genes are sequences that are not genes at all but are misinterpreted by the automated gene caller (e.g. GeneMark). Such programs are often fooled by short open reading frames that aren't really genes.

To address the last concern, you resolve to limit your attention to large genes (> 1000 nt). These are beyond mistake by gene callers. But what about the first concern? You need to know:

Are large genes that are annotated hypothetical any more likely than
other genes to reside in regions with unusual dinucleotide biases?

Strategy:

1. Find the difference in mean nucleotide bias between the two sets
    a. Create Set 1, the set of large genes annotated as hypothetical
       - Find all genes annotated as hypothetical
         (GENES-DESCRIBED-BY …)
       - Loop through these genes, collecting only those that are large
         (FOR-EACH Result option: WHEN …. COLLECT)
         (ORDER … > …)

    b. Create Set 2, the set of all other genes
       (SUBTRACT-SET …)

    c. Calculate the mean nucleotide biases of Set 1 and Set 2
       (ENTER Strep-world)
       (BIAS-OF)

2. Determine significance of the difference between the means via t-test
   (T-TEST) n.b. See Resources/Links for T-Test P calculator
3. Determine significance of the difference between the means via a simulation